Linear Regression

Starter

1. **(Review of last lesson)**

A group of students failed these marks in their Pure Maths and Statistics exams.

Student	A	B	C	D	E	F	G	H
Pure Maths	62	52	48	79	36	47	44	42
Statistics	74	66	52	71	56	73	40	57

(a) Rank each set of marks.
(b) Calculate, to 4 d.p., Spearman's rank correlation coefficient.
(c) Test, at the $5\,\%$ level the hypothesis that there is a positive association between relative performance in the two exams.

Working: (a)

Student	A	B	C	D	E	F	G	H
Pure Maths	2	3	4	1	8	5	6	7
Statistics	1	4	7	3	6	2	8	5

(b) Differences in ranks: $+1,\ -1,\ -3,\ -2,\ +2,\ +3,\ -2,\ +2$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(1^2+(-2)^2+(-2)^2+1^2+2^2+0^2+0^2)}{8(8^2-1)}$$

$$r_s = 1 - \frac{6\times 36}{512-8} \approx 0.5714$$

$$r_s = \frac{4}{7} \approx 0.5714$$

(c) '***Positive association***' suggest a $1-$tailed test.
H_0 : there is no association between the two variables
H_1 : there is a positive association between the two variables
The $5\,\%$ critical value for a $1-$tailed test with 8 values is 0.6429
Since $r_s = 0.5714 < 0.6429$ there is no evidence to reject H_0
i.e. there is no evidence to suggest Pure Maths and Statistics results have a positive association

E.g. 1 The variables x and y are known to be linearly related. Fifty pairs of experimental observations of the two variables gave these results:

$$\sum x = 402.0,\ \sum y = 83.4,\ \sum xy = 680.2,\ \sum x^2 = 3238.2,\ \sum y^2 = 384.6.$$

(a) Obtain the regression line for y on x, giving constants to 4 sf.

(b) Estimate, to 4 sf, the value of y when $x = 7.8$.

Working

(a)
$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 680.2 - \frac{402.0 \times 83.4}{50} = 9.664$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 3238.2 - \frac{402.0^2}{50} = 6.12$$

$$b = \frac{S_{xy}}{S_{xx}} \approx 1.579 \text{ (4 s.f.)}$$

$$a = \bar{y} - b\bar{x} = \frac{83.4}{50} - 1.579 \times \frac{402.0}{50} = -11.03 \text{ (4 s.f.)}$$

$$y = 1.579x - 11.03$$

(b) When $x = 7.8$, $y = 1.579 \times 7.8 - 11.03 = 1.288$

E.g. 2 A farm food supplier monitors the number of hens kept, h, against the weekly consumption of food, f kg, for a sample of 10 smallholdings. The results are below:

$$\sum h = 360,\ \sum h^2 = 17362,\ \sum f = 286,\ \sum f^2 = 10928.94 \text{ and } \sum hf = 13773.6$$

(a) State which the independent variable is.

(b) Obtain the regression equation f on h in the form $f = a + bh$, giving constants to 3 s.f..

(c) Give a practical interpretation of the gradient b.

(d) If food costs £7.50 for a 25 kg bag, estimate the weekly cost of feeding 48 hens.

Working

(a) Number of hens kept, h

(b)
$$S_{hf} = \sum hf - \frac{\sum h \sum f}{n} = 13773.6 - \frac{360 \times 286}{10} = 3477.6$$

$$S_{hh} = \sum h^2 - \frac{\left(\sum h\right)^2}{n} = 17362 - \frac{360^2}{10} = 4402$$

$$b = \frac{S_{hf}}{S_{hh}} = \frac{3477.6}{4402} = 0.790 \text{ (3 s.f.)}$$

$$a = \bar{f} - b\bar{h} = \frac{286}{10} - 0.790 \times \frac{360}{10} = 0.160 \text{ (3 s.f.)}$$

$$f = 0.160 + 0.790h$$

(c) The amount of extra food per extra hen

(d) When $h = 48$: $f = 0.160 + 0.790 \times 48 = 38.08$ kg of food

Weekly cost $= \frac{38.08}{25} \times$ £$7.50 =$ £11.424

The weekly cost would be £11.42

Using your calculator to find the regression line

Menu >> 6: Statistics >> 2: y=a+bx >> (Enter the data) >> AC >> OPTN >> 3: Regression Calc

Video (calculator): Special calculator function to find the equation of the regression line

E.g. 3 Find the regression line of y on x for the following data.

x	2	4	5	8	10
y	3	7	8	13	17

Working:

$a = -0.294117647$
$b = 1.705882353$
$r = 0.998024104$ ***this is the PMCC value***
So $y = 1.71x - 0.294$

E.g. 4 For a set of data the linear regression line is given by $y = 9.4x - 7.8$. Find the regression line for s on t given that $y = 2s - 5$ and $x = 1 - 6t$.

Working:

$2s - 5 = 9.4(1 - 6t) - 7.8$
$s = 3.3 - 28.2t$

E.g. 5 Consider the following situations and identify the independent and dependent variables, if there are any:

(a) A researcher wants to see if there is a connection between test mark and revision time.

(b) A park ranger wonders whether a greater number of trees in a wood increases the number of squirrels present.

(c) A farmer wants to know whether putting different fertiliser on crops will make them grow more.

(d) The speed of a migrating bird was measured at one-minute intervals.

Working:

(a) "Revision time" is the independent variable as it can be controlled and comes before the "test mark". "Test mark" depends on revision time so is the dependent variable.

(b) Neither "number of trees" not "number of squirrels" is an independent variable because the ranger cannot control them, unless she plants extra trees and controls how many trees are in each wood.

(c) "Type of fertiliser" is the independent variable and "growth of crops" is the dependent variable.

(d) "Time" is the independent variable because the time intervals are controlled by the experimenter. "Speed of the migrating bird" is the dependent variable.

Video: Linear regression
Video: Interpreting lines of best fit

Linear regression EQ

Solutions to Starter and E.g.s

Exercise

p86 5C Qu 1i, 2i, 3-6, (7-8 red)